ITEA Journal 2010; 31: 112–120

Applying the System Component and Operationally Relevant Evaluation (SCORE) Framework to Evaluate Advanced Military Technologies

Craig Schlenoff
National Institute of Standards and Technology, Gaithersburg, Maryland

The System, Component and Operationally Relevant Evaluation (SCORE) Framework has been developed at the National Institute of Standards and Technology over the past 3 years to provide formative evaluations of advanced military technologies. SCORE is a unified set of criteria and software tools for defining a performance evaluation approach for complex intelligent systems. To date, SCORE has been used to evaluate a wide range of advanced technologies, including soldier-worn sensor systems, technologies allowing real-time multimedia information sharing among soldiers in the field, two-way speech translation systems, and autonomous robotic platforms.

Key words: Emerging technologies, end-user utility, face recognition, intelligent systems, soldier-worn sensors, sound recognition, technological performance.

Designing and implementing a performance evaluation of an emerging technology to present a broad picture of technology performance in its typical operating environment is a very challenging goal. Intelligent systems tend to be complex and non-deterministic, involving numerous components that are jointly working together to accomplish an overall goal. As intelligent systems emerge and take shape, it is important to understand their capabilities and limitations. Evaluations are a means to assess both quantitative technical performance and qualitative end-user utility.

The System, Component and Operationally Relevant Evaluation (SCORE) Framework has been developed at the National Institute of Standards and Technology (NIST) over the past 3 years to provide formative evaluations of advanced military technologies that are still under development. SCORE is built around the premise that, in order to get a true picture of how a system performs in the field, it must be evaluated at the component level, the system level, the capability level, and within operationally relevant environments.

SCORE is a unified set of criteria and software tools for defining a performance evaluation approach for complex intelligent systems. It provides a comprehensive evaluation blueprint that assesses the technical performance of a system and its components through isolating and changing variables as well as capturing end-user utility of the system in realistic use-case environments.

SCORE is unique in that it is applicable to a wide range of technologies, from manufacturing to defense systems; elements of SCORE can be decoupled and customized based upon evaluation goals; it has the capability for evaluating a technology at various stages of development, from conceptual to full maturation; and it combines the results of targeted evaluations to produce an extensive picture of a system's capabilities and utility.

To date, SCORE has been used to evaluate a wide range of advanced technologies, including soldier-worn sensor systems, technologies allowing real-time multimedia information sharing among soldiers in the field, two-way speech translation systems, and autonomous robotic platforms. It has been the foundation for 10 technology evaluations involving soldiers and Marines from around the country. SCORE has been used as the basis of two Defense Advanced Research Projects Agency (DARPA) programs to evaluate advanced technologies.

This article describes the details of the SCORE Framework (including showing how it is different than other evaluation approaches), chronicles how it has evolved over the past 3 years, and explains how it is has been applied to evaluate disparate advanced technologies.

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
2010		**00-00-2010 to 00-00-2010**

4. TITLE AND SUBTITLE	5a. CONTRACT NUMBER
Applying the System Component and Operationally Relevant Evaluation (SCORE) Framework to Evaluate Advanced Military Technologies	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **National Institute of Standards and Technology,100 Bureau Drive, Stop 1070,Gaithersburg,MD,20899-1070**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			
unclassified	**unclassified**	**unclassified**	**Same as Report (SAR)**	**9**	

Overview of SCORE

Intelligent systems tend to be complex and nondeterministic, involving numerous components that are jointly working together to accomplish an overall goal. Existing approaches to measuring such systems often focus on evaluating the system as a whole or on individually evaluating some of the components under very controlled, but limited, conditions. These approaches do not comprehensively and quantitatively assess the impact of variables such as environmental variables (e.g., lighting, external distances) and system variables (e.g., processing power, memory size) on the system's overall performance. The SCORE Framework, with its comprehensive evaluation criteria and software tools, was developed to enhance the ability to quantitatively and qualitatively evaluate intelligent systems at the component level—and the system level—in both controlled and operationally relevant environments.

SCORE is built around the premise that, in order to get a comprehensive picture of how a system performs in its actual use-case environment, technical performance should be evaluated at the component and system levels (Schlenoff et al. 2006). SCORE defines three evaluation goal types, as shown in *Figure 1*:

- *Component Level Testing—Technical Performance*. This type of evaluation involves decomposing a system into components to isolate those subsystems that are critical to system operation.
- *Capability Level Testing—Technical Performance*. This type of evaluation involves decomposing a system into capabilities (where the complete system is made up of multiple capabilities). A capability can be thought of as an individual functionality, such as the ability for a sensor system to send and receive pictures or the ability for a translation to identify and translate names (discussed below).
- *Capability Level Testing—Utility Assessments*. This type of evaluation assesses the utility of an individual capability. The benefit of this evaluation type is that specific capability utility and usability to the end-user can still be addressed even when the system and user-interface are still under development.
- *System Level Testing—Technical Performance*. This evaluation type is intended to assess the system as a whole in an ideal environment where test variables can be isolated and controlled. The benefit is that tests can be performed using a combination of test variables and parameters, where relationships can be determined between system behavior and these variables and parameters based upon the technical performance analysis.
- *System Level Testing—Utility Assessments*. This class of evaluation assesses a system's utility, where utility is defined as the value the application provides to the system's end user. In addition, usability is assessed, which includes effectiveness, learnability, flexibility, and user attitude towards the system.

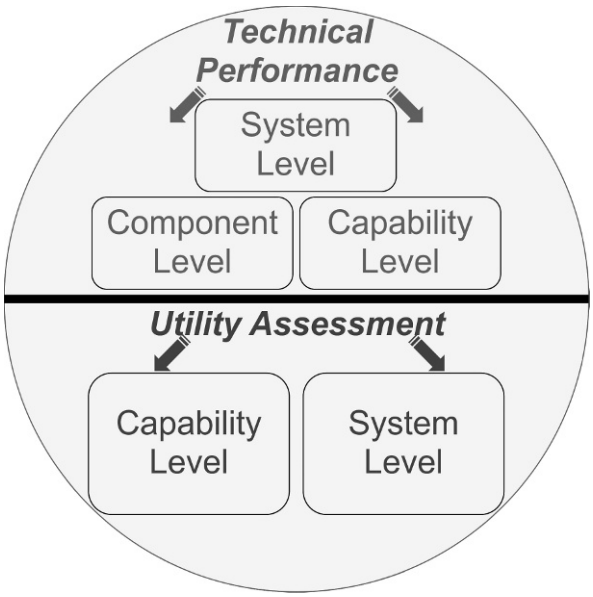

Figure 1. System, Component, and Operationally Relevant Evaluation (SCORE) Framework architecture.

Considering each of these evaluation elements, SCORE takes a tiered approach to measuring the performance of intelligent systems. At the lowest level, SCORE uses elemental tests to isolate specific components and then systematically modifies variables that could affect the performance of that component to determine the impact of those variables. Typically, this is performed for each relevant component with the system. At the next level, the overall system is tested in a highly structured environment to understand the performance of individual variables on the system as a whole. Then, individual capabilities of the system are isolated and tested for both their technical performance and their utility using task tests. Last, the technology is immersed in a longer scenario that evokes typical situations and surroundings in which the end user is asked to perform an overall mission or procedure in a highly relevant environment that stresses the overall system's capabilities. Formal surveys and semistructured interviews are used to assess the usefulness of the technology to the end user.

SCORE applied to Advanced Soldier Sensor Information Systems and Technology (ASSIST)

Overview of ASSIST

The ASSIST program is a DARPA advanced technology research and development program. The

Figure 2. A soldier interacting with one of the ASSIST technologies.

objective of the ASSIST program is to exploit soldier-worn sensors to augment a soldier's recall and reporting capability to enhance situational understanding in military operations in urban terrain environments. The National Institute of Standards and Technology Intelligent Systems Division is serving as the independent evaluation team for this program.

Technologies under test

The ASSIST program is developing a variety of soldier-worn sensors, data capture, data analysis, and information presentation technologies *(Figure 2)*. Below is a listing of five of the general data types being captured and analyzed by ASSIST technologies.

Image/video data analysis capabilities

- Object Detection/Image Classification—the ability to recognize and identify objects (e.g., identify vehicles, people, license plates) through analysis of video, imagery, and/or related data sources;
- Arabic Text Translation—the ability to detect, recognize, and translate written Arabic text (e.g., in imagery data);
- Change Detection—the ability to identify changes over time in related data sources (e.g., identify differences in imagery of the same location at different times).

Audio data analysis capabilities

- Sound Recognition/Speech Recognition—the ability to identify sound events (e.g., explosions, gunshots, vehicles) and recognize speech (e.g., keyword spotting, foreign language identification);
- Shooter Localization/Shooter Classification—the ability to identify gunshots in the environment (e.g., through analysis of audio data), including the type of weapon producing those shots, and the location of the shooter for those gunshots.

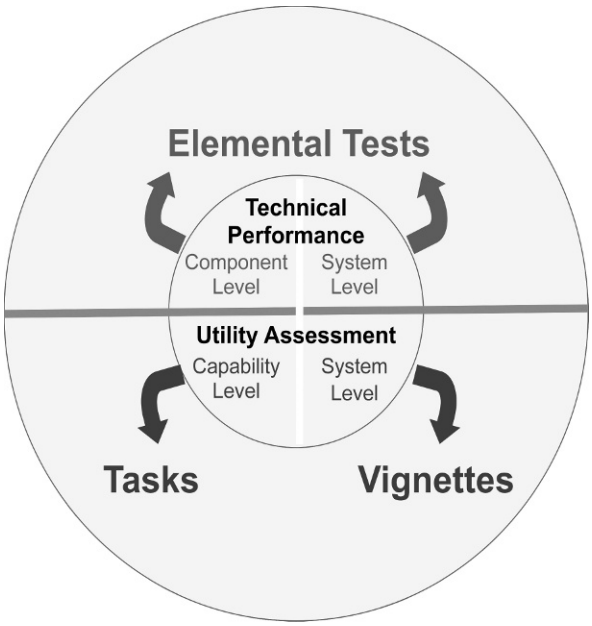

Figure 3. SCORE applied to ASSIST.

Soldier activity data analysis capabilities

- Soldier State Identification/Soldier Localization—the ability to identify a soldier's path of movement around an environment and characterize the actions taken by the soldier (e.g., running, walking, climbing stairs).

Real-time image capture and exchange

- The ability of a soldier to take a picture and send it to other soldiers in the field in real time.

Face recognition

- The ability for a soldier using a face recognition system to match the face in the picture with preloaded faces of people in a database.

Testing methodology

Elemental tests

Figure 3 shows how the SCORE Framework was applied to ASSIST. The technical performance of the ASSIST systems (both at the system and component level) were evaluated via elemental tests. In short, elemental tests were designed to measure the progressive development of ASSIST system technical capabilities. In specifying the detailed procedures for each elemental test, the independent evaluation team attempted to define evaluation strategies that would provide a reasonable level of difficulty for system and soldier performance.

Elemental tests were developed to test ASSIST technologies in an "ideal" environment and allowed focused examination of specific system components. While these tests did not immerse the technologies in realistic military scenarios, they afforded the ability to modify certain variables in a controlled fashion to assess the impact of those variables on technology performance in a military operations in urban terrain environment. Examples of elemental tests developed for ASSIST include the following:

- A shooter localization test that determined the ability for a system to identify gunshots, the type of weapon producing those shots, and the source of those gunshots in an environment with some obstructions and minimal background noise. A "zero line" and four firing lines (≈50 m, ≈100 m, ≈200 m, ≈300 m) were marked on the firing range. Simple wooden-walled structures (single story and two story) with windows were constructed at the firing lines and in the sensor region to simulate the buildings and obstructions that would be found in a military operations in urban terrain environment. Variables included
  - shooter positioning relative to walls at the firing line (within a window, next to a wall, from a clearing), and
  - obstructions between the firing line and sensor field (positions obstructed by walls that could occlude a bullet's muzzle blast and/or shockwave from a subset of the sensors).
- A soldier state/localization test that determined the ASSIST systems' ability to localize a soldier in indoor and outdoor environments, and to characterize the motion of the soldier (e.g., running, walking, going inside a building, going up stairs, lying down). Different tests exposed the system to different levels of difficulty, including inside versus outside, open versus global positioning system (GPS)-hampered locations, changes in elevation, etc. One hundred one waypoints were marked with 2-cm accuracy using differential GPS and surveying equipment. There were 42 indoor points across two different levels of buildings. There were 59 outdoor points, about 20 of which were placed next to walls and buildings, thus making it difficult to pick up a GPS signal.
- An object classification test evaluated the capabilities of the ASSIST systems to classify imagery based on the presence of various objects (e.g., people, vehicles, weapons) and states (outdoors and indoors). Approximately 50 waypoints were marked with 2-cm accuracy using differential GPS and surveying equipment. The waypoints

included a range of indoor, outdoor, ground-level, and upper-story locations (including positions in front of doorways, windows, and other building features). These waypoints were used to mark the locations from which imagery would be captured by the ASSIST-wearer, and the locations of additional objects to be placed in the environment. Imagery was collected from 25 viewpoints, each of which had multiple viewpoints to capture data from different orientations.

- A sound recognition test evaluated the ASSIST system's ability to detect certain sounds in the environment. Scripted sounds included the following:
  - firing of blank rounds from one of three weapons: M240, M4, M107;
  - a person standing next to the ASSIST wearer speaking one of ten text phrases that incorporated some combination of the keywords;
  - a person in the environment speaking foreign languages; and
  - vehicles either accelerating or decelerating past the ASSIST wearer.

  There were seven runs, each of increasing complexity. During the early runs, there was little or no ambient noise, the ASSIST wearer was stationary, there were no overlapping sounds, and most of the sounds in the environment occurred fairly close to the ASSIST wearer. During the later runs, there was a lot of ambient noise, the ASSIST wearer was moving, there were overlapping sounds, and the sounds in the environment were moving to and from distances further from the ASSIST wearer. The last two runs in the evaluation incorporated the ASSIST wearer being in confined and indoor locations. Ground truth locations of the ASSIST wearer and the sounds in the environment were measured based upon known points in the environment.
- An Arabic text elemental test evaluated the ASSIST system's ability to detect, recognize, and translate Arabic signs. Three signs were placed in the environment at marked positions, so that sets of images could be taken at known angles and distances from the signs. The first sign contained hand-printed characters, while the other two had machine-printed characters. The elemental test had the following three parts:

  1. *Sign Detection*. The signs were used to evaluate the ability of the system to extract text regions from signs.
  2. *Text Extraction*. The regions extracted from the signs were processed and the results evaluated. In addition, pictures of text were submitted to the optical character recognition program. The output Arabic characters and words were compared with those on the signs. The fonts and point sizes of the text were controlled and were limited to those that the optical character recognition system could handle.
  3. *Text Translation*. A set of Arabic words and sentences was input to the translation system in its preferred format and the resulting translations evaluated.

Utility tests

The utility of the system was assessed via vignette tests. Vignettes tests were designed and have previously been used to assess the value of ASSIST systems in (a) infantry squad reporting of critical information, events, and intelligence encountered during a mission, and (b) intelligence officer/intelligence operations. These tests engaged soldiers in realistic, albeit short, missions, where the ASSIST technologies were used as they conducted the missions.

One example of a vignette scenario mimicked a presence patrol. The presence patrol included leaving a forward operating base to patrol a local village, make the military presence known, and collect intelligence on the village and/or villagers before returning to the forward operating base. In this vignette, the soldiers were instructed to conduct a presence patrol in the market area of the village, and then conduct a deliberate search of the factory area.

Another vignette focused on collecting intelligence about an improvised explosive device explosion that had occurred overnight. The soldiers were instructed to gather detailed information about the improvised explosive device event. Upon completion of that mission, they were to conduct a presence patrol in the market and factory areas of the village, while attempting to identify and/or detain several "gray list" and "black list" individuals.

After these vignettes were completed, the intelligence officer was tasked with gathering data he would use to produce an intelligence report on the state of the village with respect to the upcoming election, including any related violence or unrest.

Task tests

The utility of specific capabilities of the system were evaluated using task tests. Task tests were short (10–15 minute) missions that focused on evaluating very specific capabilities of a system (e.g., the ability to take

Figure 4. An example of a task test for a checkpoint.

Figure 5. Example TRANSTAC system.

a picture and send it in real time, the ability to mark an area on a map while in the field). The missions were designed to allow the user to make heavy use of those capabilities. Specific task tests included the following:

- Street Observation and Interaction—This task was developed to specifically test real-time image sharing across multiple ASSIST systems.
- Presence Patrol—This task was designed to evaluate personnel tracking, GPS positioning, and map annotation capabilities.
- Insurgent Surveillance—This test was created to assess the capability of image and map transfer between the laptop-based systems and ground-based wearable ASSIST technologies.
- Insurgent Surveillance and Ambush—This task was created to test the ASSIST technology's ability to automatically generate significant actions based on activities in the environment.
- Base/Entry Checkpoint—This task was developed to test the face recognition/matching system's ability to capture images in the field and present matches in real time on the system-wearers personnel interface *(Figure 4)*.

Addressing each one of the SCORE Framework elements with respect to the task tests greatly enhanced the effectiveness of this series of evaluations at the most recent ASSIST events. Comprehensive utility assessments were collected from the task tests, which enabled the evaluation team to produce an extensive picture of the current state of the ASSIST technologies when combined with the elemental and vignette test data.

SCORE applied to the spoken language communication and translation system for tactical use (TRANSTAC)

Overview of TRANSTAC

TRANSTAC is another DARPA advanced technology and research program whose goal is to demonstrate capabilities to rapidly develop and field free-form, two-way speech-to-speech translation systems enabling English and Arabic speakers to communicate with one another in real-world tactical situations where an interpreter is unavailable (Weiss et al. 2008). Several prototype systems have been developed under this program for numerous military applications including force protection and medical screening. The technology has been demonstrated on personal digital assistant and laptop platforms. NIST was asked to assess the usability of the overall translation system and to individually assess each component of the system (the speech recognition, the machine translation, and the text-to-speech).

All of the TRANSTAC systems work fundamentally the same. Either English speech or an audio file was fed into the system. Automatic speech recognition was run over the speech to recognize what was said and generate a text file of the speech. That text file was then translated to another language using machine translation technology. The resulting text file was then spoken to the Arabic speaker using text-to-speech technology. This same process was then reversed when the Arabic speaker spoke *(Figure 5)*.

Testing methodology

Technical performance of the individual components of the TRANSTAC system was performed using offline tests (represented by the red arrow in *Figure 6*). Both technical and utility performance of the entire system was performed using lab-based evaluations of a laptop-based system (represented by the gray arrows in *Figure 6*) and more field-friendly utility systems (represented by the green arrows in *Figure 6*). Utility evaluations were also performed out in the field with the field-friendly systems (represented by the blue arrow in *Figure 6*). Last, the specific capabilities of the TRANSTAC systems (such as their ability to

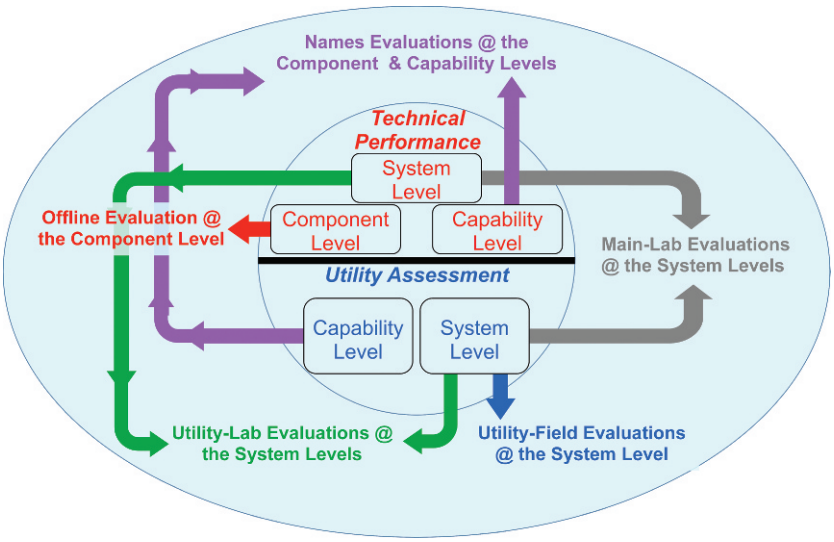

Figure 6. SCORE applied to TRANSTAC.

recognize proper names) were tested both for their technical capability and their utility (represented by the purple arrows in *Figure 6*). Each of these tests is discussed in detail below.

Offline evaluation

The offline evaluation was designed to test the TRANSTAC systems with exactly the same set of data, so comparison among the systems would truly be "apples-to-apples." Identical speech utterances, both English and Arabic, were fed into each research team's system. These utterances were collected from audio recordings from data gathering events. First, an audio file was fed into each system to test the systems' speech-to-text (S to T) capabilities. Then a text format was fed in to test their systems' text-to-text (T to T) capabilities. Since the system outputs include translated text and speech, metrics were extracted through comparison of the system outputs to ground truth. A range of metrics including low-level concept transfer and automated metrics were able to be extracted from the offline outputs (Sanders et al. 2008).

The use-case scenarios under which the utterances (both speech and text) were generated stem from the supporting data collections (and their respective scenarios) that take place months in advance of the evaluation. Appropriate scenarios were chosen based on interviews with relevant military personnel and Arabic speakers to determine the representative use-cases in which this type of technology would be most beneficial. The data collections brought together English and Arabic speakers to role-play through the numerous data collection scenarios that produced 10- to 20-minute data collection dialogues. Each of the audio dialogues were transcribed and translated where a majority of the data was provided to the research team to train their systems while the remainder was held back to create the evaluation scenarios.

Approximately 3,200 of these held-back utterances were used for the offline evaluation set. Analysis of the offline evaluation focused on component level analysis of the TRANSTAC systems using automated metrics and human judgments. The following metrics were used to analyze the offline data:

- Automatic Speech Recognition
  - Word Error Rate (automated metric)
- Automatic Speech Recognition and Machine Translation together
  - METEOR, BLEU (automated metrics)
  - Fine-grained concept transfer, performed by bilingual human judges (counting how many content words were translated properly)
  - Likert judgment at utterance level, performed by bilingual human judges
- Text-to-Speech Evaluation
  - Word Error Rate (automated metric)
  - Likert judgment performed by bilingual human judges

Live lab-based evaluation

Twenty-one structured 10-minute scenarios were created for the live lab evaluation at three stations. Structured scenarios provided a set of questions to the English speaker that they needed to find answers to. The Arabic speaker was given the answers to those

Figure 7. Environment for the Field Evaluations.

questions in paragraph format. A dialogue occurred between the two speakers and the number of answers that the English speaker was able to obtain was noted. In addition, questionnaires were provided to the English and Arabic speakers to gauge their perception of the TRANSTAC systems.

The lab evaluations were designed to test the TRANSTAC systems in an idealistic environment, with no background noise and with the participants being stationary. The TRANSTAC systems were placed on a table as opposed to being worn by the speakers. This idealistic environment gave the evaluation team and the developers an idea of the best performance of the systems at this stage in their development.

For the structured scenarios, the following metrics were used to analyze the data:

- A count of high-level concepts found out by the English speaker in response to the questions they were given. Also counted were the number of times the English question came across, number of times the answer came across, and the number of times the English speaker reported that they got the answer.
- Analysis of the questionnaire performed by English and the Arabic speakers after each scenario they participated.

Live field-based evaluation

The purpose of the field evaluations was to test the TRANSTAC systems in a more realistic environment. These tests focused on how well the systems could be carried, how easy they were to use, how well they handled wind and background noise, etc. The English-speakers carried the TRANSTAC system, and the speakers were mobile during the evaluation. Dialogues were open-ended but had to stay in the topic area of the scenarios. Each scenario lasted approximately 15 minutes. Two field scenarios were developed to gauge the utility of the TRANSTAC systems. The scenarios were performed outdoors with the English speakers wearing combat gear (body armor, helmet, gloves, etc.). They carried a "utility version" of the TRANSTAC systems while performing the scenarios. Various props were provided in the environment to make the scenario more realistic *(Figure 7)*.

Following the scenarios, the English speakers filled out questionnaires and participated in interview sessions with the evaluation team. This field exercise only looked at the utility of the system, not its technical capability. Because the utility version of the systems were on different hardware platforms than the systems used in the rest of the evaluation, the evaluation team conducted a small "utility technical evaluation" in the lab environment, which evaluated these utility versions to the laptop version by running three of the same scenarios used in the main evaluation again on the utility platforms.

Capability evaluation

The goal of the capability evaluation was to isolate specific functionalities of a system and test its performance with scenarios that were tailored to stress that functionality. For TRANSTAC, the evaluation team focused on the ability for the TRANSTAC system to identify and convey proper names in a dialogue. Three unique, names-laden scenarios were created as scripted dialogues and recorded by unique speaker-pairs. These dialogues were created such that there was at least one proper name in each Arabic utterance. These recorded data were used to create the offline names evaluation.

The offline names evaluation was run similar to that of the other offline evaluations. Specific utterances were selected and fed directly into the TRANSTAC systems. However, the measures and metrics from this test focused on how the systems specifically handled the translations of the proper names.

The live names evaluation was run in a different manner than that of the live lab evaluation. The speakers were provided with the scripted names scenarios and instructed to read them verbatim. After hearing the English utterance, the Arabic speaker responded with their scripted utterance, which was spoken into the TRANSTAC system. If the English speaker was able to understand the name that was communicated, they noted that and moved on to the next utterance. If the English speaker was unable to

ascertain a name from the TRANSTAC output, then they were able to rephrase their utterance in any manner they saw fit. Likewise, the Arabic speaker, upon hearing the English speaker rephrase their utterance, could rephrase theirs accordingly to convey the desired name. The output of this evaluation produced both technical performance and utility assessment data. This took the form of measuring the number of names successfully transferred per unit time and collecting survey responses from the end users regarding their specific names interactions.

Conclusion/future work

SCORE has proven to be an invaluable evaluation design tool of the NIST Evaluation Team and was the backbone of 10 (six for ASSIST and four for TRANSTAC) successful evaluations. It is expected to play a critical role in the remaining ASSIST and TRANSTAC evaluations.

The SCORE Framework is applicable to domains beyond emerging military technologies and those solely dealing with intelligent systems. Personnel at NIST are applying the SCORE Framework to the virtual manufacturing automation competition and the virtual RoboRescue competition (within the domain of urban search and rescue). Their intent is to develop elemental tests and vignette scenarios to test complex system capabilities and their component functions. The framework has proven to be highly adaptable and capable of meeting most any evaluation requirement.[1] ❑

CRAIG SCHLENOFF received his bachelor of science degree in mechanical engineering from the University of Maryland, College Park, and his master's degree in mechanical engineering from Rensselaer Polytechnic Institute. He is the acting group leader of the Knowledge Systems Group in the Intelligent Systems Division at the National Institute of Standards and Technology (NIST). His research interests include performance evaluation techniques, primarily applied to military advanced technologies and autonomous systems. He is currently managing multiple million-dollar efforts focusing on performance evaluation of military advanced technologies in support of multiple Defense Advanced Research Projects Agency (DARPA) programs. This work was awarded the Department of Commerce Bronze Award, the highest award that NIST can bestow. E-mail: craig.schlenoff@nist.gov

Endnotes

[1]Disclaimer: Certain commercial products and software are identified in this article in order to explain our research. Such identification does not imply recommendation or endorsement by NIST, nor does it imply that the products and software identified are necessarily the best available for the purpose.

References

Sanders, G., S. Bronsart, S. Condon, C. Schlenoff. 2008. Odds of successful transfer of low-level concepts: A key metric for bidirectional speech-to-speech machine translation in DARPA's TRANSTAC program. In *Proceedings of LREC 2008*, May 28–30, Marrakech, Morocco. Paris, France: European Language Rescources Association.

Schlenoff, C., M. Steves, B. A. Weiss, M. Shneier, and A. Virts. 2006. Applying SCORE to field-based performance evaluations of soldier worn sensor technologies. *Journal of Field Robotics* 24(8–9): 671–698.

Weiss, B. A., C. Schlenoff, G. Sanders, M. Steves, S. Condon, J. Phillips, and D. Parvaz. 2008. Performance evaluation of speech translation systems. In *Proceedings of the 6th edition of the Language Resources and Evaluation Conference*, May 28–30, Marrakech, Morocco. Paris, France: European Language Resources Association.

Acknowledgments

The authors would like to acknowledge the DARPA ASSIST/TRANSTAC program manager, Dr. Mari Maeda, and the members of the NIST independent evaluation team for their continued support.